A Lexicon Based Offline Arabic Handwritten Recognition Using Naive Bayesian Classifier with Gaussian Distribution

Ahlam Hasan Albashiti
Supervisor: Hashim Tamimi
Master of Informatics
College of Graduate Studies
Palestine Polytechnic University

Introduction

Handwriting recognition is used most often to describe the ability of a computer to translate human writing into text. It can be either ”online” or ”offline” according to the way in which the text is input to the computer.

This paper is restricted to the offline handwritten recognition for the Arabic script.

Naive Bayesian Classifier with Gaussian Distribution is used to recognize words , also a lexicon is used to validate the results and give the most similar word to the tested word if it doesn't exist.

Arabic Language

Arabic script is a difficult language because of the overlapping between letters, each letter consists of one or more shapes according to its position in the word, in addition the existence of dots that change the meaning of the word. See figure 1.

Letter Name	Possible shapes			
	alone	end	middle	beginning
Alef	ا	ـا		
Ba'a	ب	ـب	ـبـ	بـ
Ta'a	ت	ـت	ـتـ	تـ
Tha'a	ث	ـث	ـثـ	ثـ
Jeem	ج	ـج	ـجـ	جـ
Ha'a	ح	ـح	ـحـ	حـ
Kha'a	خ	ـخ	ـخـ	خـ
Dal	د	ـد		
Thal	ذ	ـذ		
Raa	ر	ـر		
Zai	ز	ـز		
Seen	س	ـس	ـسـ	سـ
Sheen	ش	ـش	ـشـ	شـ
Sad	ص	ـص	ـصـ	صـ
Dad	ض	ـض	ـضـ	ضـ
TTa	ط	ـط	ـطـ	طـ
ThTha	ظ	ـظ	ـظـ	ظـ
Ein	ع	ـع	ـعـ	عـ
Gein	غ	ـغ	ـغـ	غـ
Faa	ف	ـف	ـفـ	فـ
Qaf	ق	ـق	ـقـ	قـ
Kaf	ك	ـك	ـكـ	كـ

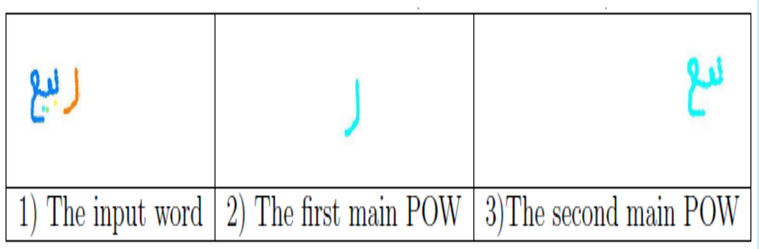
Figure 1 : shapes of characters according to the position

Project Objectives

Building a system for Arabic handwritten recognition using Naive Bayesian Classifier with Gaussian Distribution based on lexicon.

Proposed project

The main phases are :
Preprocessing : it includes image binarization, morphological operation, and word segmentation using connected component method see figure 2.



1) The input word	2) The first main POW	3)The second main POW

Figure 2 : Segmentation of the word into POWs

Feature extraction: features for each Part Of Word (POW) is extracted using connected component to for a feature vector of length 16 .

Classification: Naïve Bayesian classifier with Gaussian Distribution is used to map any feature vector to its corresponding class.

Recognized Word: In order to get the possible probability for the tested word, we make a combinations between all POWs then we find the probability for each combination.

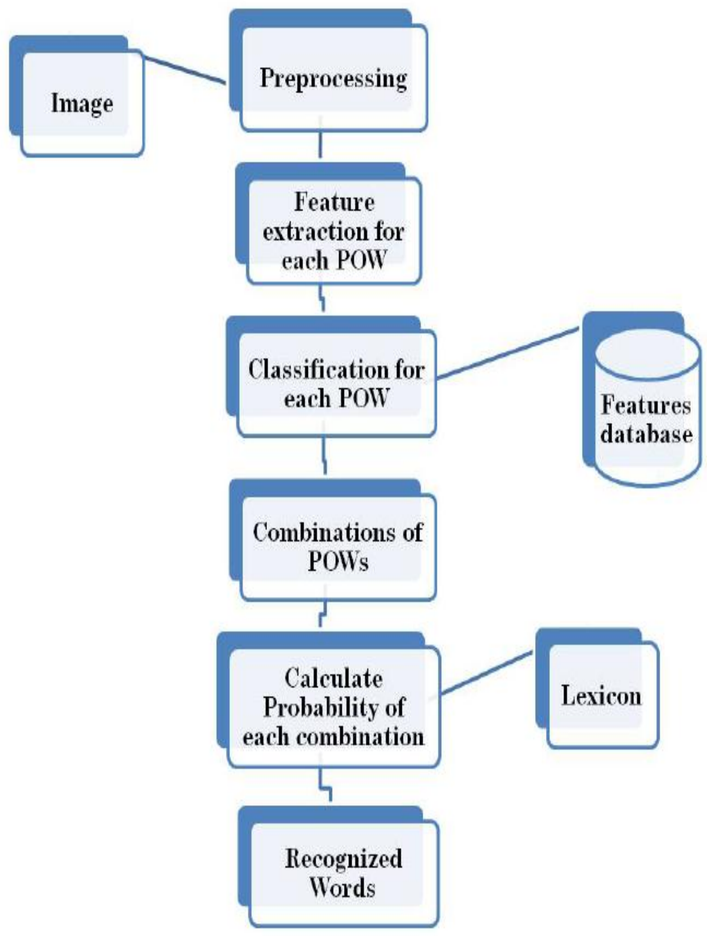


Figure 3: General block diagram for the proposed system

Results:

There is no reference data set for training and testing samples for the offline Arabic word recognition. Therefore we built our own data set for training and testing. According to the training data set we built 300 samples for 30 POW for two writers .For the testing we built 2 testing sets, first set consists of 29 words, the POW of these words are taken from the training data set, while second set test contains of POW that are not trained.

Number of posteriors	Success rate
5	0.7679
6	0.7857
7	0.7946
8	0.8214
9	0.8482
10	0.8839

Table 1 :Result Of POW for connected components features

Number of posteriors	Success rate	Recognized word
1	0. 6786	19
5	0.9286	26
8	0.9287	26

Table 2: Result Of words from the POW that are exist in the training data

Number of posteriors	Success rate	Recognized word
1	0.4286	12
3	0.7679	19
5	0.7857	22
8	0.7857	22

Table 3: Result Of words from the POW that are not exist in the training